*De novo* transcriptomes built from hundreds of eye tissues reveal hundreds of novel gene isoforms

Vinay S Swamy, Temesgen D Fufa, Robert B Hufnagel, David M McGaughey

## Introduction

       The transcriptome is defined as the set of distinct RNA transcripts expressed in a population of identical cells. During transcription several RNA processing steps modify immature RNA and drive the formation of multiple, distinct isoforms for most genes. For example, the human Gencode release 28 contains 97,713 protein coding transcripts across 20,306 genes (1) . RNA processing broadly describes a variety of biological mechanisms and includes alternative promoter usage, alternative splicing, RNA editing, and alternative polyadenylation. The full biological impact of gene isoforms has not been fully elucidated, but multiple studies have shown that gene isoforms can have distinct and critical functions in biological processes like development(2), cell differentiation(3), and cell migration(4). Alternative usage of isoforms has also been implicated in multiple diseases including cancer(5), cardiovascular disease (6), Alzheimer’s Disease(7) and diabetic retinopathy(8).(didnt read these papers so need to double check them)

       Some of the first methods using RNA-seq to find novel gene isoforms focused on identifying novel exon-exon junctions and novel exon boundaries based on RNA-seq coverage. (9) More recently, several groups have developed specialized tools to use RNA-seq to reconstruct the whole transcriptome of a biological samples, dubbed *de novo* transcriptome construction (10),(3),(11).

*De novo* transcriptome construction uses short (~100bp) RNA-seq reads to reconstruct full-length mRNA transcripts. However, a large number of samples are necessary to overcome the noise and short read lengths of this type of data. Because of increasingly inexpensive sequencing, data sets of the necessary size are now available. For example, one of the most comprehensive *de novo* transcriptome projects to date has been CHESS, which used the GTEx data set to construct *de novo* transcriptomes in over 9000 RNA-seq samples from 49 distinct location of the body to create a comprehensive annotation of mRNA transcripts across the human body. (12),(13) However, as the GTEx data set lacks any ocular tissues, the CHESS database is an incomplete annotation of the human transcriptome.

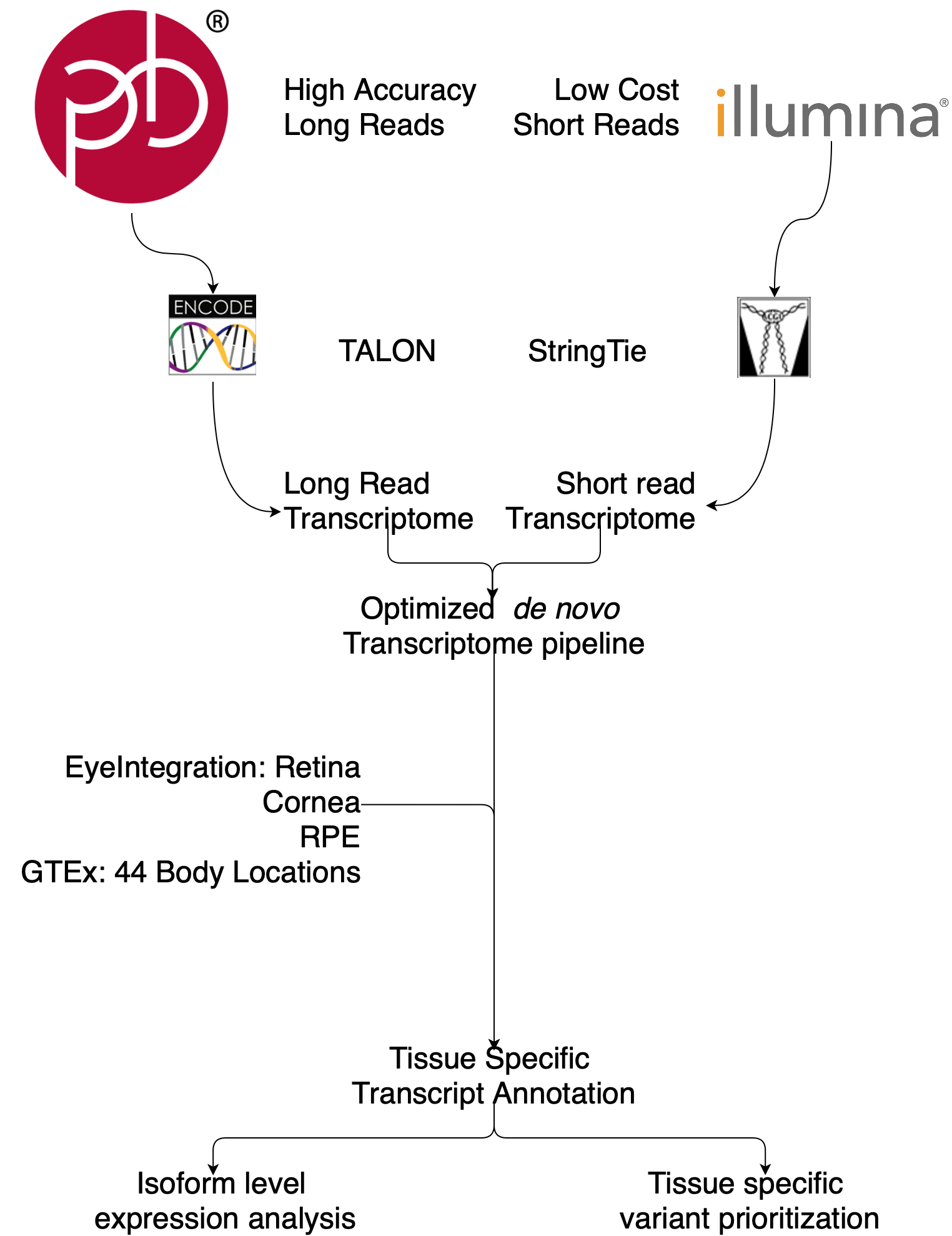
       Despite the increasing number of tools for transcriptome construction there has been no gold standard with which to evaluate precision and sensitivity of *de novo* transcriptome construction on real (not simulated) biological data. Long read sequencing technologies provide a potential solution to this problem as long read sequencing can capture full length transcripts and thus can be used to identify a fuller range of gene isoforms. While long reads have historically been considered inaccurate, the new PacBio Sequel II system sequences long reads as accurately as short read based sequencing (14).

       We propose that long read based transcriptomes can serve as a ground truth for evaluating short-read base transcriptomes, and in this study we use PacBio long read RNA sequencing to inform the construction of short read transcriptomes. We generated PacBio long read RNA-seq data from a stem cell derived retinal pigmented epithelium (RPE) cell line along with matched Illumina short read RNA-seq. Using the two sources of RNA-seq data we design a rigorous Stringtie-based *de novo* transcriptome pipeline that maximizes the concordance between short and long read *de novo* transcriptomes.

       We apply this pipeline using a previously published data set containing 368 ocular tissue samples compiled from mining publicly available sequencing data (15). We use this pipeline to build transcriptomes in three major ocular subtissues: The cornea, retina, and the RPE, using RNA-seq data from both adult and fetal tissues to create a high-quality pan-eye transcriptome. In addition to our ocular samples, we used a subset of the GTEx data set to construct transcriptomes for 49 other locations across the body to facilitate comparisons between transcriptomes across the body.

       We use our gold-standard informed pan eye *de novo* transcriptome to reveal hundreds of novel gene isoforms as well as several novel transcribed genomic loci in the eye and analyze their potential impact on ocular biology and disease. We provide our *de novo* transcriptomes as a resource to other researchers through an R package

## Methods



(Supplemental) Figure 1

Supplemental Figure 1. Workflow for De novo Transcriptome construction and analysis.

## Generation of PacBio long read RNA sequencing data and Illumina short read RNA sequencing

       Human induced pluripotent stem cells (iPSCs) were differentiated into mature RPE following the culturing protocol in Blenkinsop et al(16). RNA was isolated from mature RPE 40 days post differentiation and used for <illumina library prep> and <pacbio library prep>

## Code availability and software versions.

       To improve reproducibility, we wrote all code used to generate both the data and figures for this paper as Snakemake pipelines. (17) All code (and versions) used for this project is publicly available in the following github repositories: <https://github.com/vinay-swamy/ocular_transcriptomes_pipeline> (main pipeline),

<https://github.com/vinay-swamy/ocular_transcriptomes_longread_analysis> (long read analysis pipeline), <https://github.com/vinay-swamy/ocular_transcriptomes_paper> (figures and tables for this paper), <https://github.com/vinay-swamy/ocular_transcriptomes_shiny> ( webapp) .

WE ALSO NEED TO GIVE A GIT TAG FOR EACH TO INDICATE THE “RELEASE” VERSION WITH THIS PAPER (ALSO DEPOSIT IN ZENODO).

## Computational Analyses

All computational analyses performed in this project are run using multiple Snakemake workflows. Each Snakefile contains the exact parameters for all tools and scripts used in each analysis. All Snakefiles are included as supplementary data.(supplementary data files 1-4)

## Analysis of Long Read Data

      PacBio hifi reads were processed into full length, non-chimeric (FLNC) reads using the Pacbio SMRT link software  [VERSION][]{.comment-end id=“3”}. The existing ENCODE longread RNA-seq pipeline (<https://github.com/ENCODE-DCC/long-read-rna-pipeline>) was rewritten as a Snakemake workflow as follows. Transcripts were aligned to the human genome using minimap2(18), using an alignment index built on the gencode v28 primary human genome. Sequencing errors in aligned long reads were corrected using TranscriptClean (19), using default paramters. Splice junctions for TranscriptClean were obtained using the TranscrtiptClean accessory script “get\_SJs\_from\_gtf.py” using the gencode v28 comprehensive annotation as the input. A list of common variants to avoid correcting were obtained from the ENCODE portal (<https://www.encodeproject.org/files/ENCFF911UGW/>). The long read transcriptome was generated with TALON (20). A TALON database was generated using talon\_initialize\_database command, with all default parameters, except for the “–5P” and “–3p” parameters. These parameters represent the maximum distance between close 5’ start and 3’ ends of similar transcript to merge, and were both set to 100 to match parameters used in later tools. Annotation in gtf format was generated using the talon\_create\_GTF command, and transcript abundance values were generated using the talon\_abundance command.

## Analysis of short read RPE data

      We aligned each sample to the Gencode release 28 hg38 assembly using the genomic aligner STAR and sorted the resulting BAM files using samtools sort. (1),(21),(22). For each sorted BAM file, we constructed a per sample base transcriptome using StringTie with the Gencode V28 comprehensive annotation as a guiding annotation (1),(11). All sample transcriptomes were merged with the long read transcriptome using gffcompare(23) with default parameters. We note that the default values for the distance to merge similar 5’ starts and 3 ends of transcripts in gffcompare is the same to what we chose for TALON. We define the metric construction accuracy, used to evaluate short read transcriptome construction as the following:

## Construction of tissue specific transcriptomes.

       We used studies with healthy, unperturbed RNA-seq samples from 52 distinct subtissue regions of the body, downloaded and performed quality control the pertinent sequencing data from the sequence read archive (SRA) using methods from our previous work(15). We constructed a transcriptome for each sample, and merged samples together to create 52 subtissue specific transcriptomes. For each tissue-specific transcriptome, we removed transcripts that had an average expression less than 1 Transcripts Per Million (TPM) across all samples of the same tissue type. All tissue specific transcriptomes were merged to form a single unified GTF annotation file to ensure transcript identifiers were the same across tissues. We merged all ocular tissue transcriptomes to generate a separate pan-eye transcriptome.

## Tissue specific transcriptome quantification

       For each resulting tissue specific transcriptome, we extracted transcript sequences using the tool gffread (23) and used these sequences to build a tissue-specific quantification index using the index mode of the alignment free quantification tool Salmon. (24) For each sample, we quantified transcript expression using the quant mode of salmon, using a sample’s respective tissue specific quantification index. We similarly quantified all ocular samples using the pan-eye transcriptome and the Gencode v28 reference transcriptome.

## Annotation of novel exons

       Analysis of novel transcripts was done using a custom Rscript “annotate\_and\_make\_tissue\_gtfs.R” . First, a comprehensive set of distinct, annotated exons was generated by mergeing exon annotation from gencode, ensembl, UCSC, and refseq. We then defined a novel exon as any exon within our trancriptome that does not excatly match the chromosome, start, end and strand of an annotated exon. Novels exons were classified as followed. Exons were split into 3 catagories: First, last, and middle exons. We extracted all annotated exon start and stop sites from our set of previously annotated exons. Novel middle exons that had an annotated start but an unannotated end were catagorized as a novel alternative 3’ end exons and similarly novel middle exons with an unannoated start but annotated end were catagorized as a novel 5’ start exons. Novel middle exons whose start and end both matched annotated exon start and ends are considered retained introns. Novel middle exons whose start and end both did not match annotated starts and ends are considered fully novel exons. We then classified novel first and last exons. Novel first exons are first exons whose start is not in the set of annotared exon starts, and novel last exons are terminal exons whose end is not in the set of annotated exon ends.

## Validation of DNTX with phylop, CAGE data, and polyA signals

       PhyloP (25) scores for the phylop 20 way multi species alignment was downloaded from UCSC’s FTP server on October 16th, 2019 and converted from bigWig format to bed format using the wig2bed tool in BEDOPs (26) . The average score per exon in both the gencode and DNTX annotation was calulated by intersecting exon locations with phylop scores and then averaging the per base score for each exon, using the intersect and groupby tools from the bedtools suite, respectively. Significant difference in mean phylop score was tested with a wilcox rank sum test.        CAGE peaks (27) were download from the FANTOM FTP server (<https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/CAGE_peaks/hg38_fair+new_CAGE_peaks_phase1and2.bed.gz>) on June 15th 2020. Transcriptional start sites (TSS) were extracted from gencode and DNTX annotations; TSS is defined as the start of the first exon of a transcript. Distance to CAGE peaks was calculated using the closest tool in the bedtools suite. Significant difference in mean distance to CAGE peak between DNTX and gencode annotation was tested with a wilcox rank sum test.        Polyadenylation signal annotations were downloaded from the polyA site atlas (28) (<https://polyasite.unibas.ch/download/atlas/2.0/GRCh38.96/atlas.clusters.2.0.GRCh38.96.bed.gz>) on June 15th 2020. Transcriptional end sites(TES) were extracted from gencode and DNTX annotations; TES is defined as the end of the terminal exon of a transcript. Distance to polyA signal was calculated using the closest tool in the bedtools (29) suite. Significant difference in mean distance to polyA signal was tested with a wilcox rank sum test.

## Identification of protein coding novel transcripts.

       We identified protein coding transcripts in our unified transcriptome using the TransDecoder suite (10). We extracted transcript sequences using the util script “gtf\_genome\_to\_cdna\_fasta.pl” and used TransDecoder to find a single best open reading frame from each transcript. We then used the “agat\_sp\_add\_start\_stop.pl” scripts from the AGAT tool (<https://github.com/NBISweden/AGAT/>) to identify start and stop codons for each open reading frame. Transcripts with no detectable ORF or missing a start or stop codon were labelled as noncoding. Additionally, novel isoforms whose predicted amino acid sequence’s length was either 200 amino acids shorter than the shortest annotated isoform for that gene were marked as noncoding. Similarly, novel isoforms whose predicted amino acid sequence’s length was 200 aa greater than the longest annotated isoform were also marked as noncoding.

## Prediction of novel loci function

       Protein coding novel loci were compared to the uniprot protein sequence data base using blastp(30). Blastp results were integrated with hmmer, which identified protein families and domains associated with each novel loci.

## Analysis of novel isoforms in eye tissues.

An Upset (31) plot was generated using the ComplexUpset package(<https://github.com/krassowski/complex-upset>). Fraction Isoform Usage(FIU) for each calcualted each transcript t associated with a parent gene g using the following formula: . Raincloud plots of FIU were generated using the R\_Rainclouds package. (32)

## Prediction of Variant impact using *de novo* transcriptomes.

Clinvar variants were downloaded from the clinvar database in VCF format and ClinVar metadata in tab delmited format were downloded on April 19th 2020. The VCF of variants was used as the input variants for the Variant Effect Predictor from Ensembl (33) . Tissue specific variant priority was determined using gtfs corresponding to each tissue specific transcriptome as input annotation. Variant predictions were also generated using the gencode v28 comprhensive transcript annotation. Variants marked as “Uncertain Significance” in the tab delimited clinvar file were used in down stream analysis. The following query was used to search for variants associated with ocular disease: ‘macula|retin|leber|cone|cornea|bardet|ocular|optic|joubert’

## Analysis of fetal retina RNA-seq data.

       RNA-seq samples from Mellough et al. were downloaded from the SRA using methods from a previous study (15) . Samples were quantified using salmon with a quantification index generated using our fetal retina *de novo* transcriptome. We removed samples deemed as outliers by first performing principal component analysis of transcript level expression data, calculating the center of all data using the first two principal compenents, and removing the 5 samples furthest away from the center. The remaining samples were normalized using calcNormFactors from the edgeR (34) R package and converted to weights using the voom function from the limma R package. (35) Differential expression was modeled using the lmFit function using developmental time point as the model design and tested for signficant change in expression using the Ebayes function from limma. Gene Set enrichment was testing using the clusterprofileR package. (36) Heatmaps were generated using the ComplexHeatmap package (37) .

## Computing Resources

       All computation was performed on the National Institutes of Health high performance compute system Biowulf (hpc.nih.gov).

## Figures and Tables

       All statistical analyses, figures and tables in this paper were generated using the R programming language. (38) A full list of packages and versions can be found in supplementary file session\_info.txt

# Results

## Long Read Pacbio RNA sequencing guides *de novo* transcriptome construction

In order to evaluate the accuracy of short read transcriptome construction, we first generated PacBio long read RNA-seq data and Illumina short read RNA-seq data from a stem cell derived RPE cell line. These cell lines were cultured using a highly optimized protocol, and thus should have minimal biological variation (16). We used this sequencing data to construct a long read transcriptome and a short read transcriptome.(methods) In our long read transcriptome we find 1163239 distinct transcripts, and in our short read transcriptome 366888 distinct transcripts

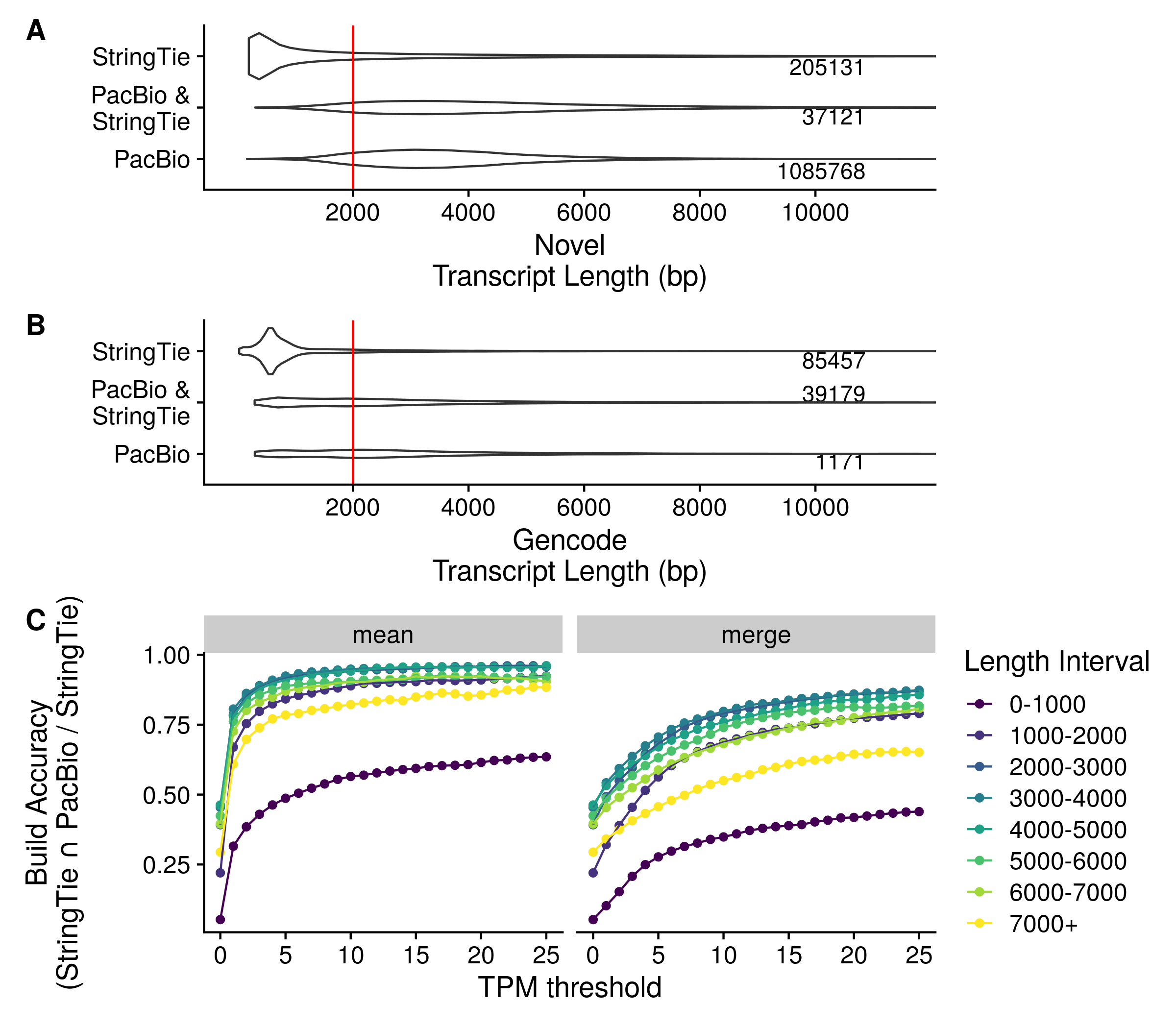


Figure 1. Transcript lengths are substantially longer in the long-read based approach. A,B) Intersection of novel and annotated transcript lengths between Pacbio (long read) and Stringtie (short read) transcriptomes. The total number of constructed transcripts is given in the text to the right of the violin plot. C) Short read construction accuracy stratified by transcript length at different TPM based transcript exclusion thresholds.

    In our initial comparison between short and and long read transcriptomes, we see a low transcriptome construction accuracy (see Methods) of 0.208. When examining the transcript lengths of each build we see that the two methods show very different transcript length distributions for both novel and previously annotated transcripts, with the short read build comprised mostly of smaller transcripts (Fig 1A). As the PacBio data was generated using two libraries for 2000 bp and >3000 bp (need to double check this), we expected to see enrichment for longer transcripts in the pacbio data set. (Supplemental Fig 2) To assess accuracy relative to transcript length, we group transcripts by length in 1000 bp intervals, and compare accuracy between each group. We found the accuracy significantly improves for transcripts longer than 2000 bp. The construction accuracy is 0.426 and 0.137 for transcripts above and below 2000 bp, respectively.(Fig 1B)     We experimented with various methods to remove spurious transcripts to improve construction accuracy. We first removed transcripts that were not expressed at 1 TPM in at least one sample as outlined in the stringtie’s recommended protocol. (39) This improved construction accuracy to 0.475 for transcripts longer than 2000bp and 0.212 for transcripts shorter than 2000bp. As this accuracy was still fairly low, we tried a different filtering scheme. We found the simplest approach with high performance was to retain transcripts that had an average TPM above a specific threshold(Fig 1C). In our downstream pipeline we keep transcripts that have at least an average of 1 TPM across all samples of the same subtissue type as this threshold achieved a build accuracy of 0.772 for transcripts longer than 2000Bp and retained 48470 transcripts within this short read RPE dataset.

## A rigorous analysis pipeline finds thousands of novel gene isoforms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tissue | Source | Samples | Studies | Transcriptome Count |
| Retina | Adult | 105 | 8 | 49714 |
| RPE | Fetal | 49 | 7 | 49967 |
| Cornea | Adult | 43 | 6 | 51469 |
| Retina | Fetal | 89 | 6 | 66255 |
| RPE | Adult | 48 | 4 | 32012 |
| Cornea | Fetal | 6 | 2 | 59408 |

Table 1. Ocular sample dataset overview and transcriptome count. Transcriptome count is defined as the number of unique transcripts expressed in a given tissue type

      We built transcriptomes from 368 published, publicly available ocular tissue RNA-seq samples using an efficient snakemake pipeline. We include both adult and fetal tissue from cornea, retina, and RPE tissues mined from 29 different studies (Table 1). Our fetal tissues consist of both human fetal tissues and human induced pluripotent stem cell (iPSC) derived tissue, as stem cell derived tissue has been showed to closely resemble fetal tissue (40) . To more accurately determine tissue specificity of novel ocular transcripts, we supplemented our publicly collated normal (non-disease, perturbation) ocular data set with 877 samples across 46 body locations from the GTEx project and constructed transcriptomes for each of these body locations (12) . We refer to each distinct body location as a subtissue here after.       After initial construction of transcriptomes, we found 183442 previously annotated transcripts and 6241675 novel transcripts detected in at least one of our 1245 samples. We define novel as any region of the human genome that has not been previously annotated within the Gencode, Ensembl, UCSC, and Refseq annotation databases. (1) , (41) , (42) After using the filtering methods desribed above, we merged all tissue specific transcriptomes into a single final transcriptome which contains 252983 distinct transcripts with 87592 previously annotated and 165391 novel transcripts, and includes 114.94673 megabases of previously unannotated genomic sequence. (Table 1) We refer to the final transcriptome as the DNTX annotation hereafter.       We split novel transcripts into two categories: novel isoforms, which are novel variations of known genes, and novel loci, which are previously unreported, entirely novel regions of transcribed sequence.(Fig 2B) Novel isoforms are further classified by the novelty of its encoded protein: an isoform with novel open reading frame, a novel isoform with a known ORF, and isoforms with no ORF as noncoding isoforms.(Fig 2A) The number of distinct ORFs is significantly less than the number of transcripts, with 41700 previously annotated ORFs and 43130 novel ORFs across all tissues. Across all tissues there is an average of 10392.85 novel isoforms and 3579.19 novel ORFs.

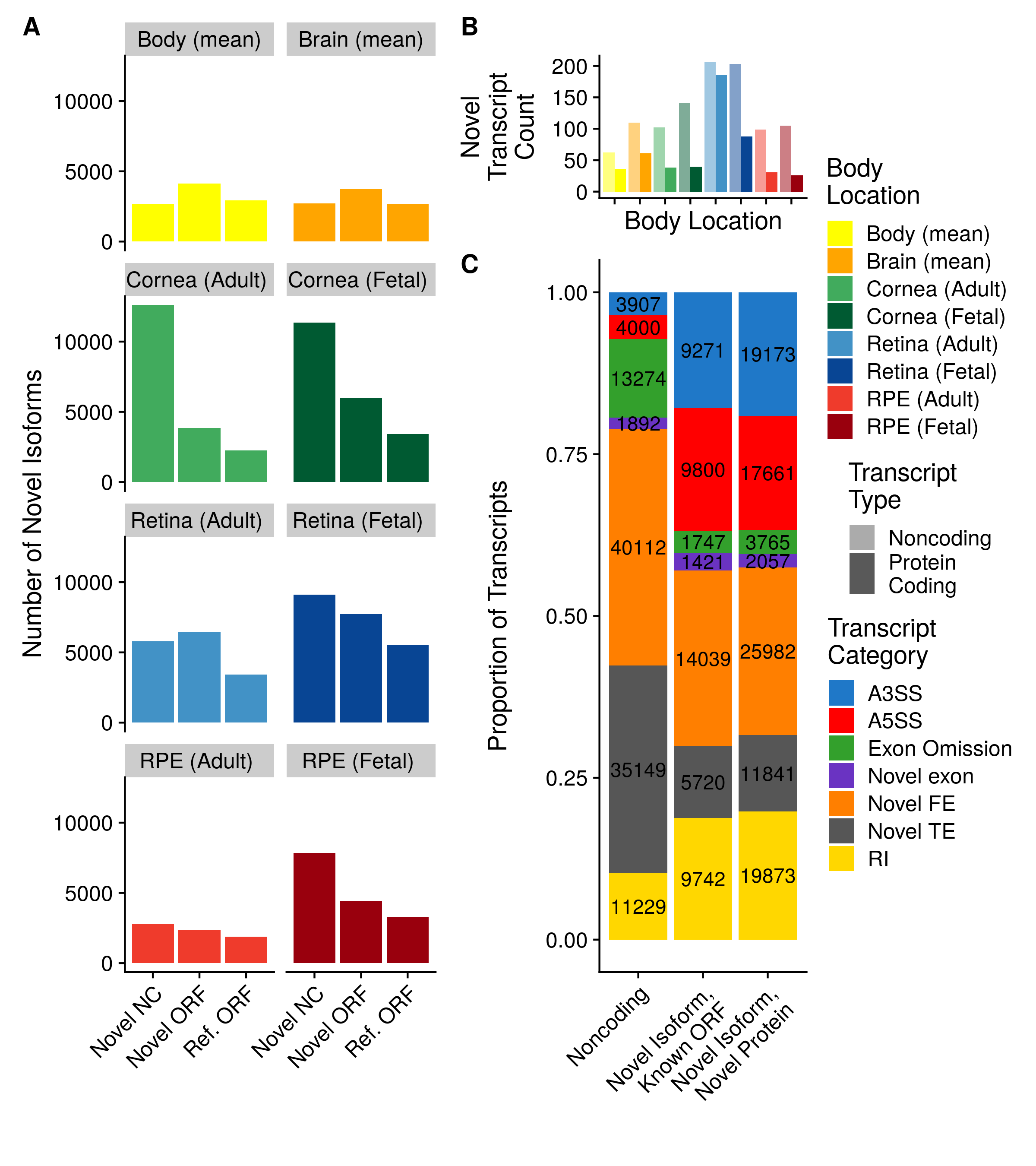


Figure 2. Overview of Novel Isoforms. A. Number of novel gene isoforms, grouped by transcript type. Brain and body represent an average of 13 and 34 distinct subtissues, respectively B. Novel protein coding and noncoding loci. Novel exon composition of novel isoforms, by isoform type labels indicate number of transcripts. C. Classification of novel exon types, stratified by novel isoform type

       Novel isoforms can occur due to an omission of a previously annotated exon, commonly referred as exon skipping or the addition of a unannotated exon which we refer to as a novel exon. We further classify novel exons by the biological process that may be driving their formations: alternative promoter usage driving the addition of novel first exons (NFE) (43), alternative polyadenylation driving the addition of novel terminal exons (NTE) (44) , and alternative splicing driving the formation of all novel exons that are not the first or last exon (45) . We further classify alternatively spliced exons into their commonly seen patterns, alternative 5’ splice site (A5SS), alternative 3’ splice site (A3SS), and retained introns (RI). Exons whose entire sequence was unannotated and is not a retained intron are fully novel exons. We note that all three of these mechanisms can lead to exon skipping, so for simplicity we group all novel isoforms resulting from exon skipping together. (considering adding a Supplemental Figure with diagrams of each group to make it easier for ppl) We found that the majority of novel exons with our dataset are novel first and last exons. We see that the majority of A5SS, A3SSs and RI exons lead to novel protein coding isoforms, whereas novel FE/TE more often lead to noncoding isoforms. (Fig 2C)

## *De novo* transcriptomes match previously published experimental data than existing annotation

      We validated our *de novo* transcriptomes using three independent datasets. First, we evaluated the conservation of our transcriptomes, as conservation has been a historic marker for function. We used PhyloP 20 way species alignment (25), a measure of conservation between species, to calculate the average conservation score for each exon our DNTX annotation, and compared that to the average conservations score for each exon in the gencode annotation. We found that on average, exons in out DNTX annoation are more conserved than exons in the gencode annotation (pvalue <2.2e-16) (Supplemental Figure 3A).       Next, as we saw an enrichment of novel first and last exons within our data set, we decided to compare the transcriptional start sites (TSS) and transcriptional end sites(TES) within our DNTX annotation to two well established annotation databases. We compared DNTX and gencode TSS’s to CAGE-seq data from the FANTOM consortium. (27) As CAGE-seq is optimized to detect the 5’ end of transcripts, we reason that it can serve as a valid ground truth set to evaluate TSS detection. (46) We calculated the absolute distance of DNTX TSS’s to CAGE peaks, and compared them to the absolute distance of gencode TSS’s to CAGE peaks. We found that on average DNTX TSS’s are closer to CAGE peaks than gencode TSS’s (pvalue <2.2e-16)(Supplemental Figure 3B). Next we evaluated TES’s using the polyA atlas, which is comprhensive annotation of polyadenylation signals generated from aggregating 3’ seq data from multiple studies. (28) As 3’-seq data is designed to accurately capture the 3’ ends of transcripts it can similarly serve as a ground truth set to evalute the accuracy of TES’s. (47) We calculated the absolute distance of DNTX TES’s to annotated polyA signalsand compared them to the absolute distance of gencode TES’s to polyA signals. We found that on average DNTX TES’s are closer to annotated polyadenylation signals than gencode TSS’s (pvalue <2.2e-16) (Supplemental Figure 3C)

## *De novo* transcriptomes reduce overall transcriptome sizes

      Our transcriptomes removed on average 76.141 % of a tissue’s base transcriptome. We define base transcriptome for a tissues as any transcript in the gencode annotation with non zero TPM in at least one sample of a given tissue type. This was a large reduction in transcriptome size and ee wanted to ensure we were not unduly throwing away data. We quantified transcript expression of our samples using Salmon, quantifying each sample twice: once using the full gencode V28 human transcript annotation, and once using its associated tissue specific transcriptome We found that despite the 76.141 % reduction in number of transcripts between the base gencode and *de novo* transcriptomes (Supplemental Figure 4A), the average salmon mapping rate increases by 2.041 % indicating that the vast majority of gene expression data is retained within our transcriptome. (Supplimental Figure 4B)

## Novel Isoforms in Ocular tissues

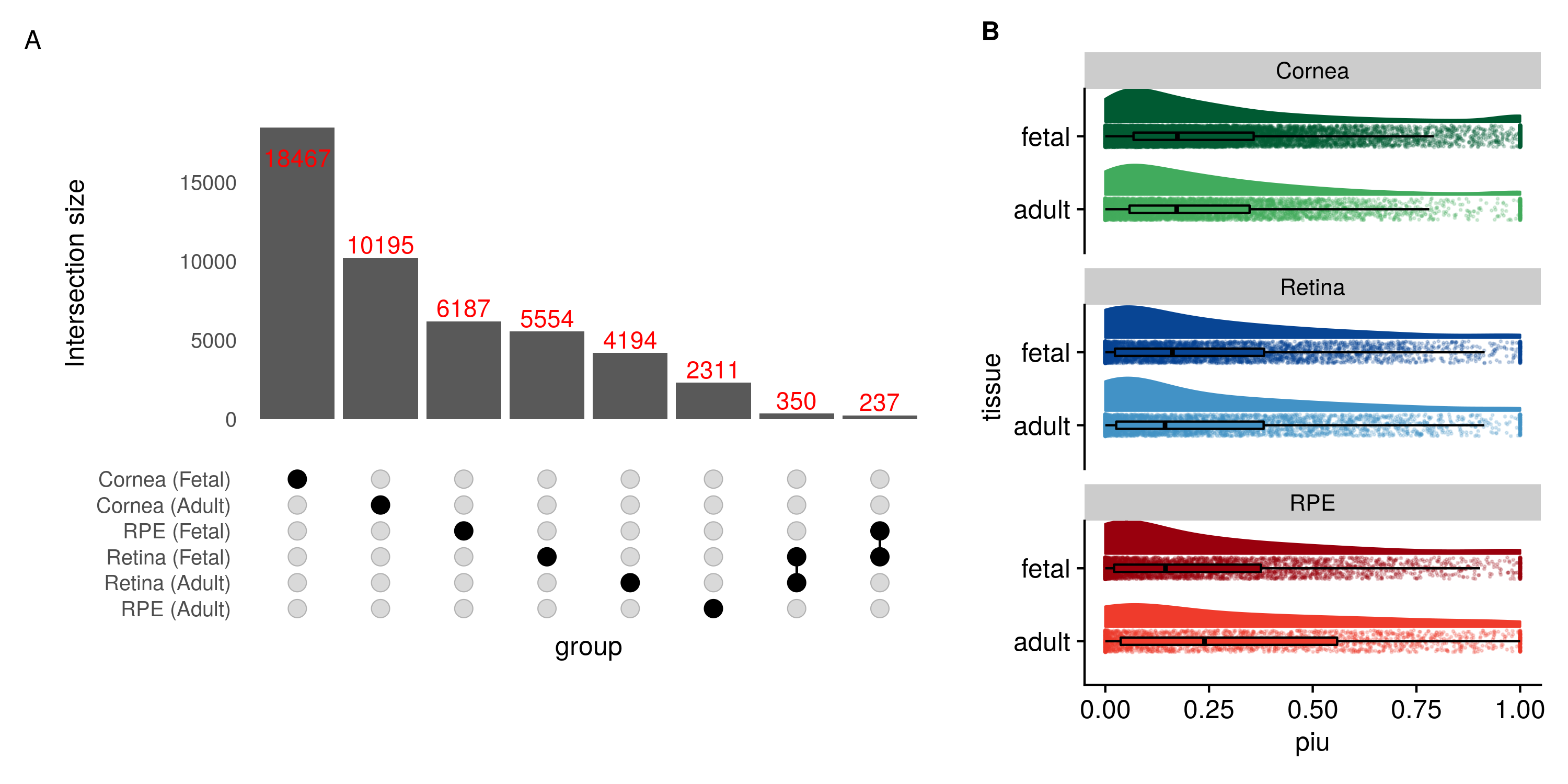


Figure 3. Analysis of novel isoform characteristics A). Set intersection of novel isoforms in ocular transcriptomes. B). Boxplots of fraction isoform usage(FIU) overlaid over FIU data points with estimated distribution of data set above each boxplot

      Next, we analyzed the novel isoforms within our ocular transcriptomes. We compared the overlap in constructed novel isoforms across ocular tissues and found that 94.99 % of novel isoforms are specific to a singular ocular subtissue (Fig 3A). For each novel isoform we then calculated fraction isoform usage (FIU), or the fraction of total gene expression a transcript contributed to its parent gene. We found that on average novel isoforms contribute to 25.79 % of their parent gene’s expression.

## Differential Usage of Gene Isoforms Occurs during Retinal Development

Mellough et al showed that alternative splicing plays a role during retinal development (48) As this is a subset of alternative isoform usage, we hypothesized that alternative isoform usage plays a role in retinal development. We use RNA-seq data from Mellough et al that we did not include in the data used to build our transcriptomes. We used our fetal retina *de novo* transciptome to quantify transcript expression and analyzed differential transcript usage (DTU) across development.

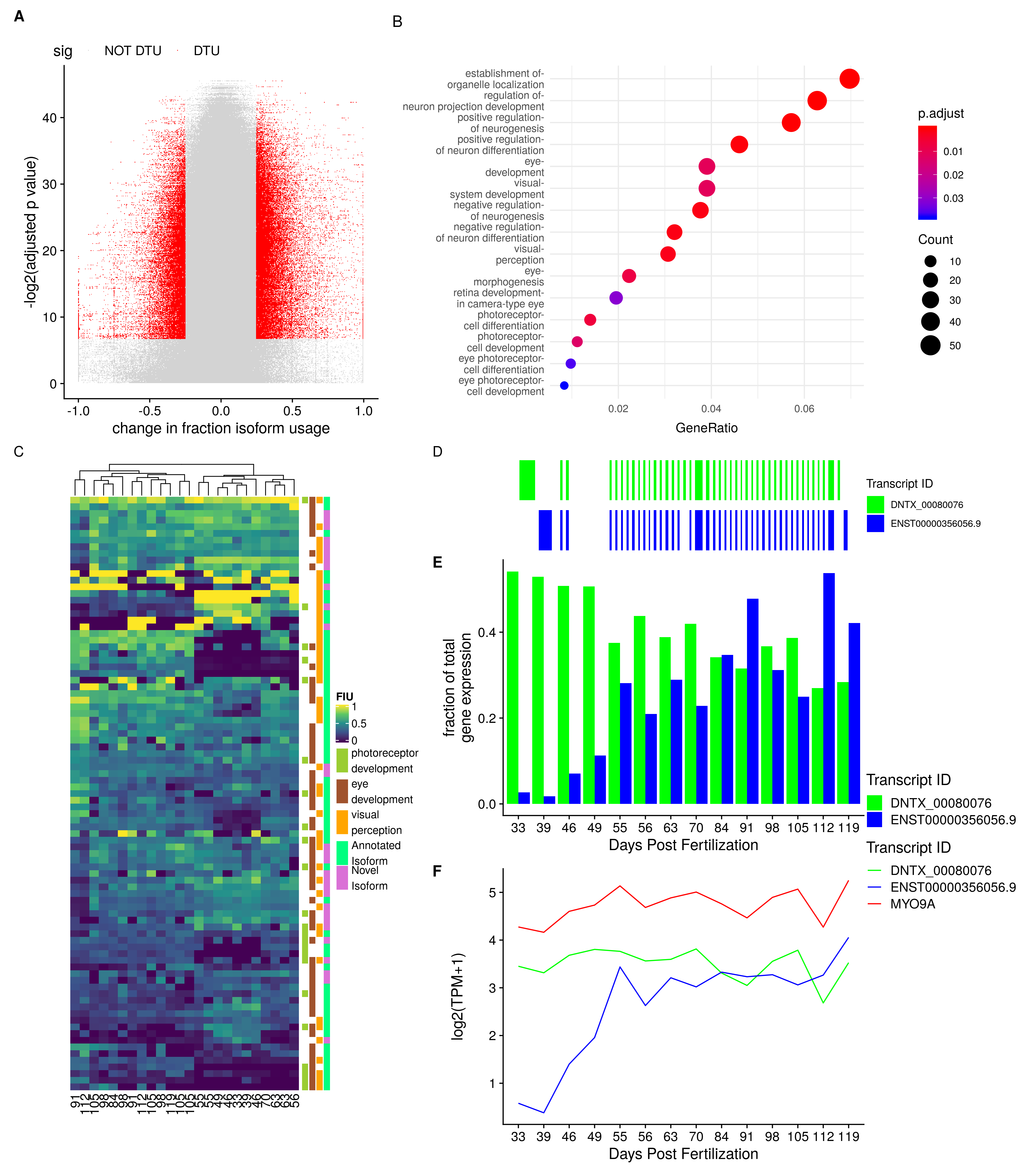


Figure 4 Differential Transcript usage during Retinal Development. A) Volcano Plot of tested transcripts B) Dot plot for gene set enrichment analysis E) Heatmap of genes with DTU associated with eye development D) Transcript models for MYO9A, a gene undergoing DTU F) change in MYO9A FIU across development F) average TPM expression of MYO9A across development

We analyzed 24 samples across 14 developmental days post fertilization and found 1717 transcripts acrosss 812 genes involved in diffferential transcript usage(DTU).(Fig 4A) We define DTU as an transcript that is differentially expressed (qvalue <.01) and has a FIU difference of .25 in at least one comparison of time points. We found that genes involved in DTU were enriched(qvalue <.05) for genes related to eye and neurological development.(Fig 4B), and that hiearchical clustering of DTU transcripts geneates an early stage and late stage cluster.(Fig 4C) One of these genes, MYO9A, is a perfect example of DTU. MYO9A is associated with the visual perception GO term and plays a role in ocular development and has been associated with ocular disease (49) . While expression of MYO9A remains relatively unchanged across development, expression of two of its associated isoforms(Fig 2D) changes across during development: a novel isoforms is highly expressed early during development, but switches to a cannonical isoform later in development.(Fig 2E,F)

## *de novo* transcriptome allow for a more precise variant prioritization.

The prediction of variant’s biological impact is a fundamental step in variant prioritization, which is a critical step in the diagnosis of genetic disease and in the interpretation GWAS results. The recent GnomAD(50) project showed that taking transcript expression in to account can improve variant prioritization. (51) We hypothesize that our *de novo* transcriptomes can provide an additional layer of refinement to expression based variant conseqeunce prediction. We use Ensembl’s Variant Effect Predictor combined with our tissue specific transcriptomes to generate a tissue specific prediction of variant impact, and reclassify variants of unknown significance (VUS) in the ClinVar database. (33),(52). We grouped different predicted consequences into a High, moderate and low impact groups and provide a full mapping of biological consequence to impact in our supplementary data.

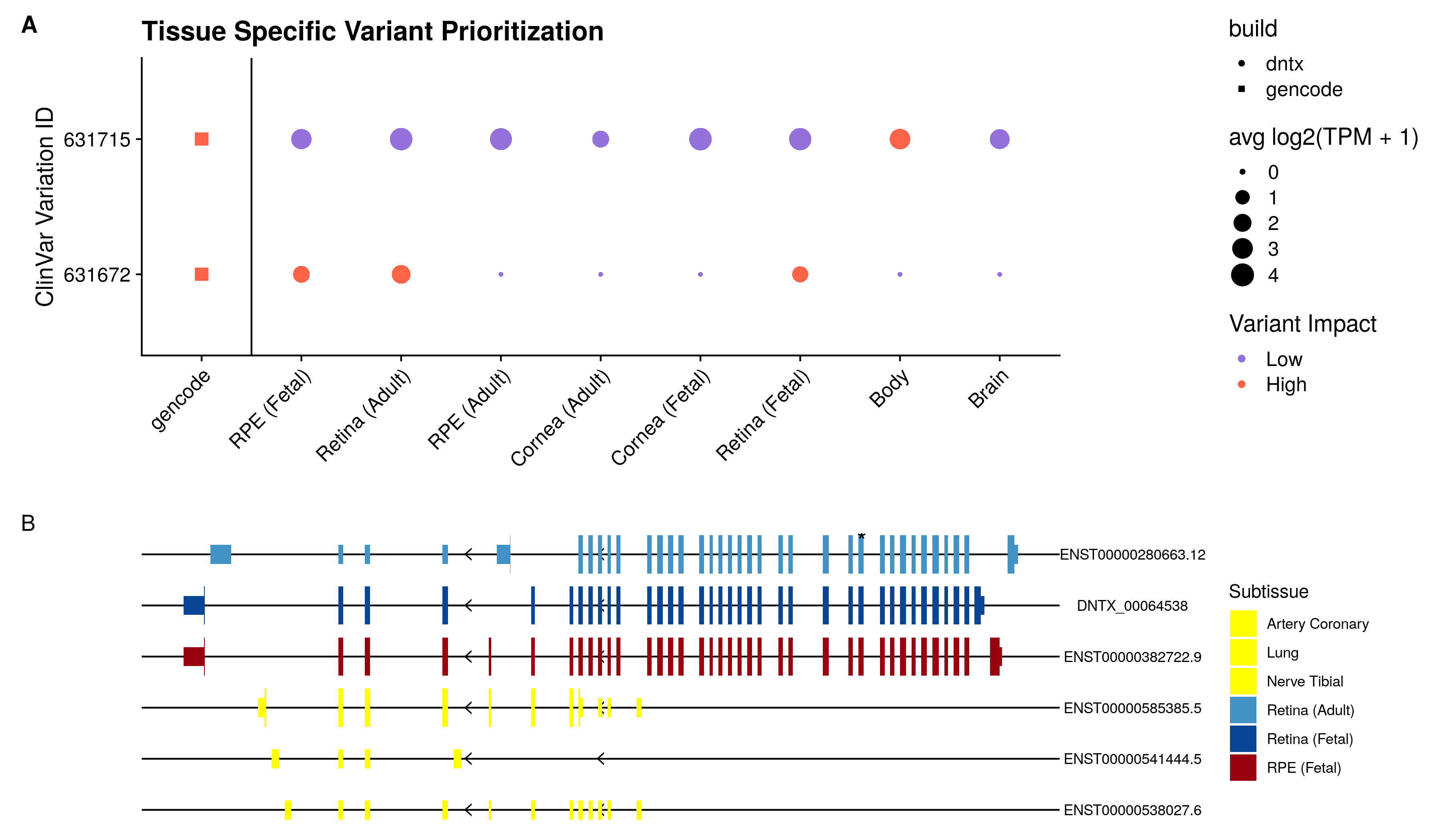


Figure 5. Tissue specific variant prioritization of two variants of unknown significance (VUS) previously associated with ocular disease. A) Variant Impact dot plot. Size corresponds to average TPM of the transcript with the highest TPM and color to impact. A full mapping of variant descriptors to impact can be found in supplemental data. b) Transcript Models of CACNA2D4 thick lines represent CDS; Asterisk represents location of variant 631715. Arrows indicate direction of transcription

We compared our tissue specific variant effect predictions to variant effect predictions generated using the gencode transcript annotation. We found that globally out of 226824 total variants examined, the priority of 118297 variants decreased in at least one tissue, and 1324 variants increased in at least one tissue. We see a larger amount of variants decrease in predicted effects as the transcript which previously lead to an effect is not expressed in a particular tissue. To demonstrate the utility of tissue specific variant prioitizaion, we focused specifically on variants previously associated with ocular disease. We examined ClinVar variant 631672, which was previously associated with retinal cone dystrophy 4, and has a high predicted impact only in ocular tissues, and not in other tissues. (Fig 5A) Variant 631672 is in the gene CACNA2D4 and in all transcripts expressed in ocular tissues this variant leads to a premature stop codon. While CACNA2D4 has other isoforms expressed in other tissues, these isoforms are not affected by variant 631672 and thus there is a higher probability that this variant may be playing a role in disease.(Fig 5B) Conversely, in clinvar variant 631715 which is associated with posterior column ataxia-retinitis pigmentosa syndrome and is within gene FLVCR2, the variant is low impact on transcripts expressed in eye tissues, but higher impact in other body tissues. This may suggest that this variant is less likely to be associated with the disease.

## A companion visualization tool enables easy use of *de novo* transcriptomes

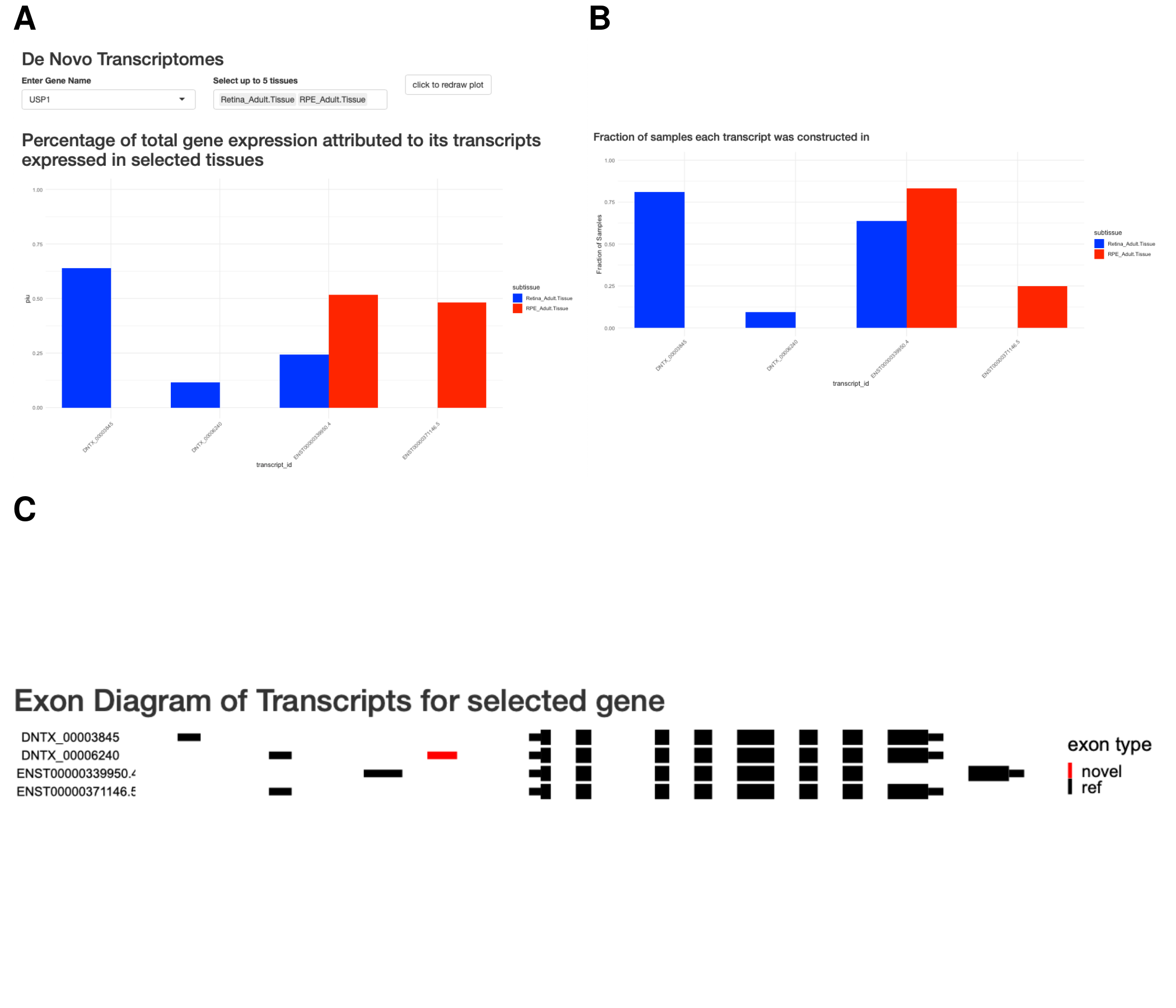


Figure 6. Screenshots from dynamic *de novo* transcriptome visualization tool. A). FIU bar plot for selected gene and tissue. B). Exon level diagram of transcript body Thicklines represent coding region of transcript. novel exons colored in red. Tooltip contains genomic location and phylop score C) Bargraph of fraction of samples within dataset each transcript was consructed in by tissue.

       To make our results easily accessible, we designed a web-app for visualizing and accessing our *de novo* transcriptomes. Users start by selecting a gene or searching for a gene by genomic location, and can choose up to 5 tissues to visualize transcript expression in. For each tissue we show the FIU for each transcript associated with a gene. (Fig 6A) We show the exon-intron structure of each transcript and mousing over exons show genomic location overlapping SNPs, and phylogentic conservation score. (Fig 6B) We additionally show a barplot of the fraction of samples in each tissue each transcript was constructed in.(Fig 6C) Users can download the *de novo* transcriptomes for selected tissues in GTF and fasta format. While visualization of direct transcript expresion is not a part of this app, it can be viewed in the eyeIntegration app (15) by selected ‘DNTX’ as the transcript annotation.

## A easy-to-use pipeline to enable future research

Finally, we package all tools used for our transcriptome pipeline within a portable docker container with a stand-alone run script. This pipeline allows other researchers to run their own samples, and generate figures and annotations similar to what is shown here

# Discussion

       Motivated by the lack of a comprehensive pan eye transcriptome we created the first comprehensive set of ocular transcriptomes. By using long read RNA-sequencing data to calibrate our short-read construction pipeline, we increase our confidence that we are creating real, biologically relevant transcriptomes. We found that concordance between long and short read based transcriptome is directly related to transcript length and transcript expression across many tissues. We saw a clear inability within this PacBio data set to accurately detect transcripts shorter than 2000Bp for both previously annotated and novel transcripts. As many of the transcripts constructed using short reads are below this threshold, long read sequencing data enriched for smaller transcript sizes would provide greater insight in future studies.

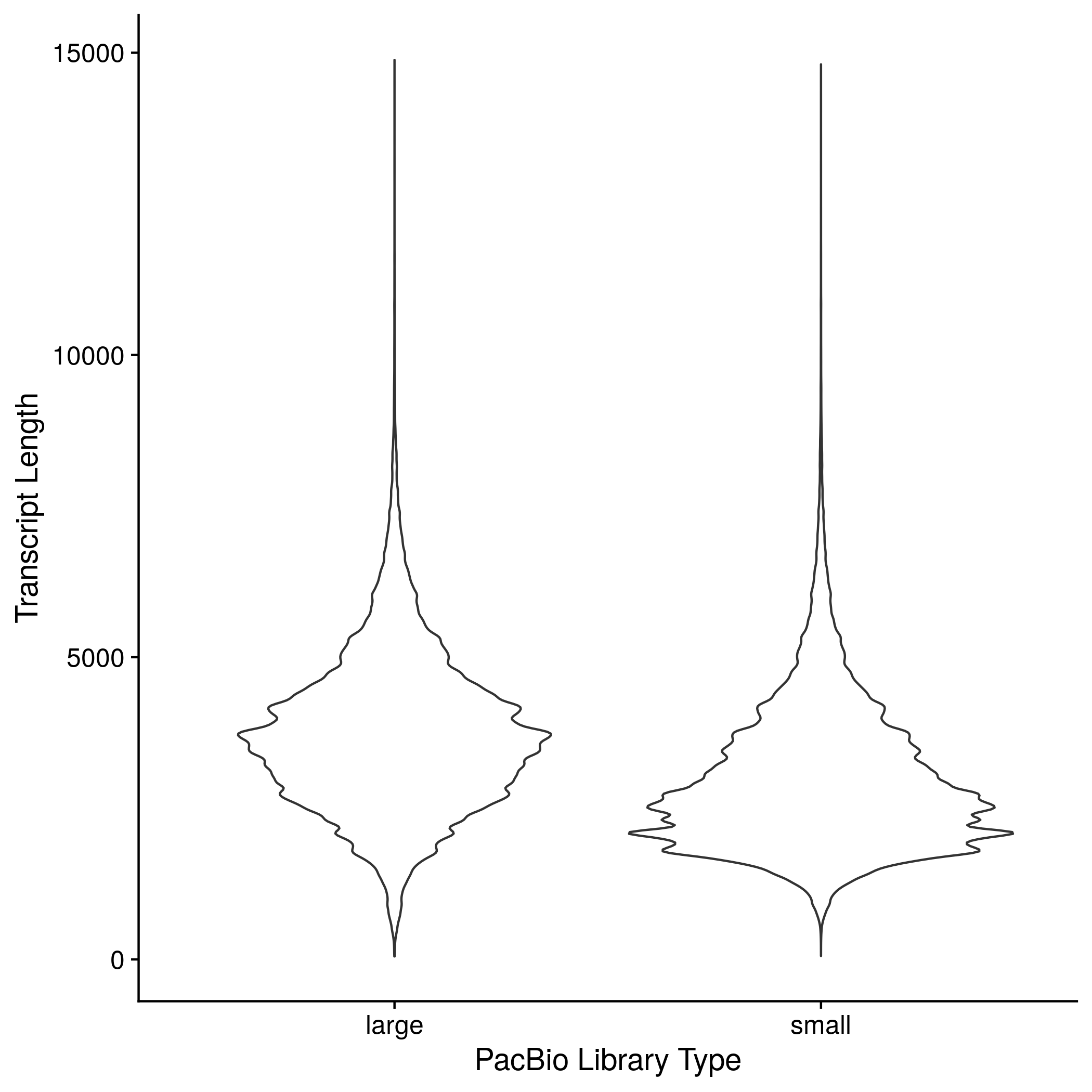
      We used a large dataset compiled from published RNA-seq data to build our ocular transcriptomes, an approach which has several key advantages. First, our large sample size allows us to combat the noisy nature of RNA-seq data. Second, as our samples come from mutliple studies and types of sample preparation because , we can be more confident that our transcriptomes accurately reflect the biology of its originating subtissue and are not a technical artifact due to preparation of the samples. Further more, we are confident that these results are accurate because our *de novo* transcritomes match existing large scale data sets and are more conserved than existing annotation. (Supplemental Figure 2)

      In each tissue we examined, we found hundereds of novel gene isoforms, many of which were novel due to novel exons. Within ocular tissues, these novel isoforms are most commonly specific to single subtissue. This makes sense as over half of the exons in our *de novo* transcirptomes are first and last exons, which have been prevously shown to significanlty contribute to the tissue specificity of gene isoforms (53) We also find that on average novel isoforms represent about 25.79 % of their parent gene’s expression. However, it is difficult to say if these are non-functional. Its is entirely possible that these isoforms are from low population cell types, as transcript annotation was previously shown to be incopmlete in rare cell types. (54) This especially makes sense in the retina which contain multiple distinct cell types, several of which contribute to 5% or less to total cell population (55) However, as we imposed a strict expression filter as part of out transcriptome pipeline, we may be removing transcripts specific to rare cell types.

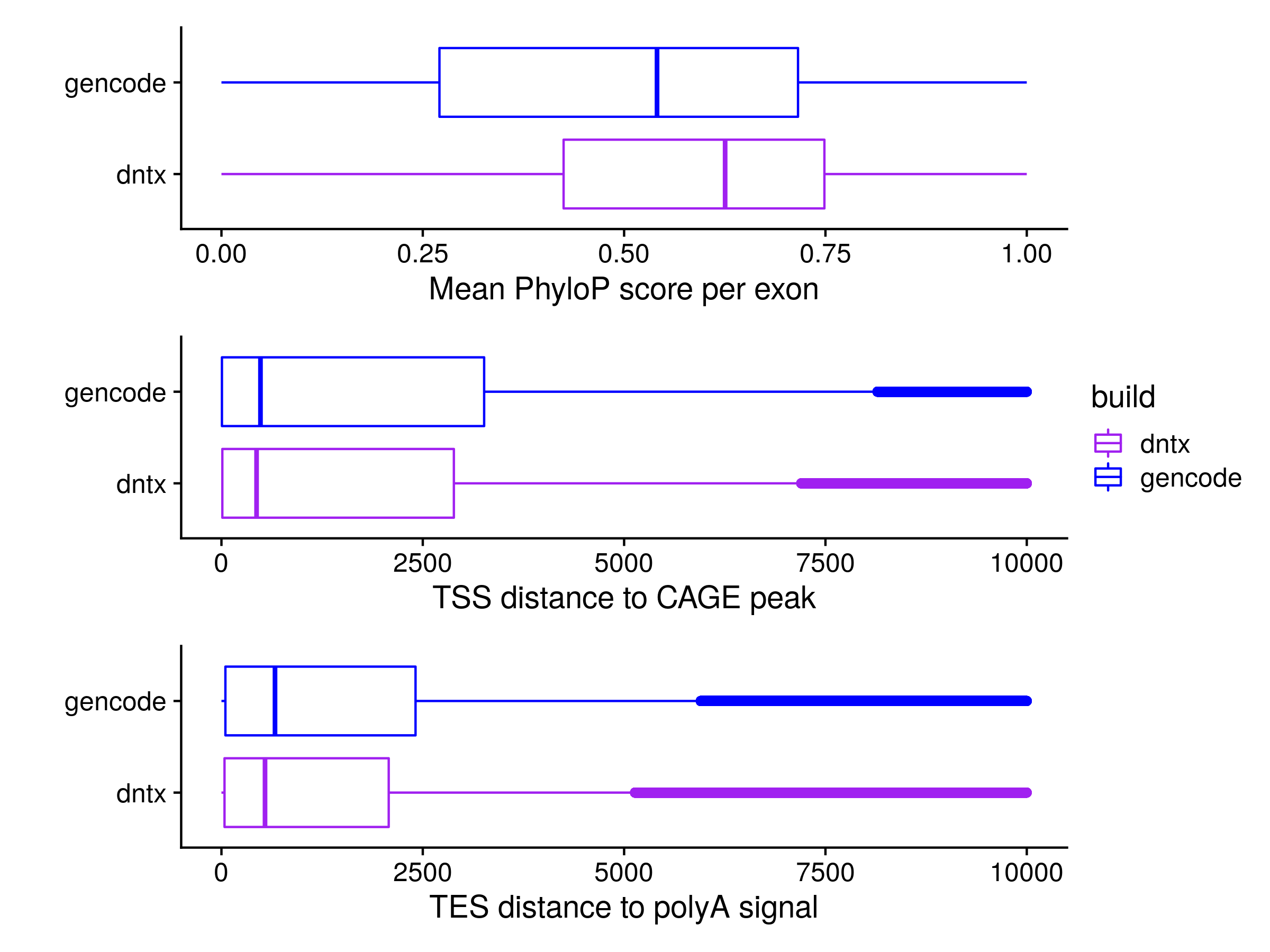
      {== We showed that our ocular transcriptomes are useful in understanding the role of gene isoforms in ocular biology and disease. We saw that multiple novel isoforms of genes associated with retinal development that exhibit significant changes in trascript expression and FIU during retinal development. We used our tissue specific transcriptomes to re prioritize variants of unknown significance, and found the tissue specific variant prioritization can increase the resolution of tissue specific variant prioritization. ==}

      This work is most useful as a starting point for other researchers; We want to make our transcriptomes easily accessible to other researchers, so we designed a webapp to visualize our transcriptomes and access tissue-specific annotation files. We wanted to provide as much information to the user so that they can draw their own conclusions about the significance of potential novel exon, and so we provide the gene model with novel exon, coding and noncoding regions marked, along with the FIU for each transcript constructed within a gene. We also provide the fraction of samples within a given subtissue type a sample was detected in, to provide a further level of evidence to the validity of constructed transcripts.

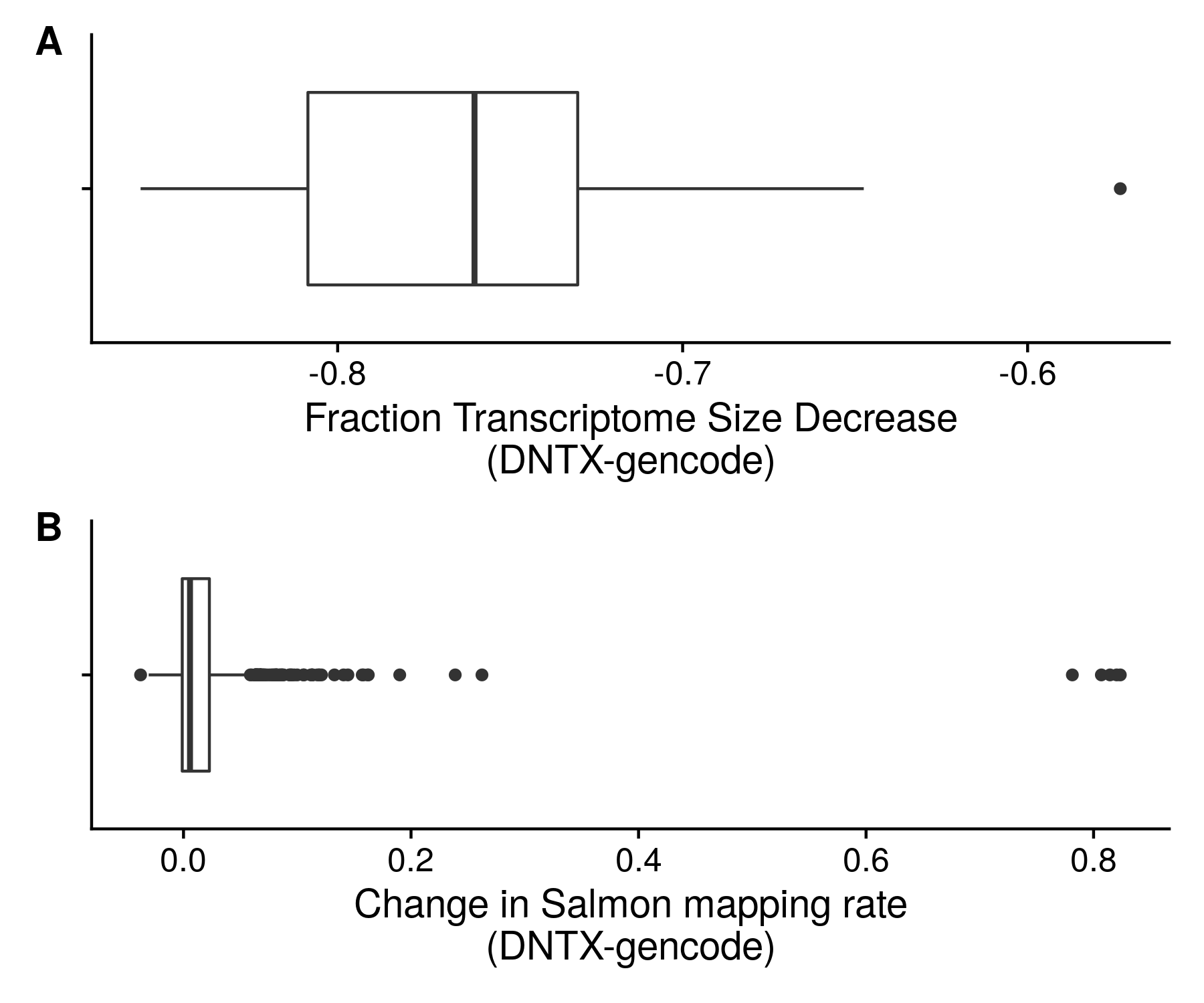
# Supplemental Figures



Suppplemental Figure 2. Distribution of PacBio long ead lengths for two library sizes



Suppplemental Figure 3. Comparison of DNTX annotation to Gencode Annotation. A) Per exon Phlop score for gencode and DNTX transcripts. B) Average distance of DNTX Transcriptional Start Sites (TSS) and Gencode TSS to CAGE-seq peaks from the FANTOM consortium. C) Average distance of DNTX Transcriptional End Sites (TES) and Gencode TES to polyadenylation signals in the PolyA site atlas.



Suppplemental Figure 4. Comparison of Salmon mapping rate change vs transcriptome size decrease.

# References

1. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J. and Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, **47**, D766–D773.

2. Dykes,I.M., Bueren,K.L. van and Scambler,P.J. (2018) HIC2 regulates isoform switching during maturation of the cardiovascular system. *Journal of Molecular and Cellular Cardiology*, **114**, 29–37.

3. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., Baren,M.J. van, Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.

4. Mitra,M., Lee,H.N. and Coller,H.A. (2020) Splicing Busts a Move: Isoform Switching Regulates Migration. *Trends in Cell Biology*, **30**, 74–85.

5. Vitting-Seerup,K. and Sandelin,A. (2017) The Landscape of Isoform Switches in Human Cancers. *Molecular Cancer Research*, **15**, 1206–1220.

6. Neagoe Ciprian, Kulke Michael, del Monte Federica, Gwathmey Judith K., de Tombe Pieter P., Hajjar Roger J. and Linke Wolfgang A. (2002) Titin Isoform Switch in Ischemic Human Heart Disease. *Circulation*, **106**, 1333–1341.

7. Mills,J.D., Nalpathamkalam,T., Jacobs,H.I.L., Janitz,C., Merico,D., Hu,P. and Janitz,M. (2013) RNA-Seq analysis of the parietal cortex in Alzheimer’s disease reveals alternatively spliced isoforms related to lipid metabolism. *Neuroscience Letters*, **536**, 90–95.

8. Perrin,R.M., Konopatskaya,O., Qiu,Y., Harper,S., Bates,D.O. and Churchill,A.J. (2005) Diabetic retinopathy is associated with a switch in splicing from anti- to pro-angiogenic isoforms of vascular endothelial growth factor. *Diabetologia*, **48**, 2422–2427.

9. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, **320**, 1344–1349.

10. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B. and Lieber,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature protocols*, **8**.

11. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.

12. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA and Biospecimen Collection Source Site—NDRI *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

13. Pertea,M., Shumate,A., Pertea,G., Varabyou,A., Breitwieser,F.P., Chang,Y.-C., Madugundu,A.K., Pandey,A. and Salzberg,S.L. (2018) CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, **19**, 208.

14. Wenger,A.M., Peluso,P., Rowell,W.J., Chang,P.-C., Hall,R.J., Concepcion,G.T., Ebler,J., Fungtammasan,A., Kolesnikov,A. and Olson,N.D. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, **37**, 1155–1162.

15. Swamy,V. and McGaughey,D. (2019) Eye in a Disk: eyeIntegration Human Pan-Eye and Body Transcriptome Database Version 1.0. *Investigative Ophthalmology & Visual Science*, **60**, 3236–3246.

16. Blenkinsop,T.A., Saini,J.S., Maminishkis,A., Bharti,K., Wan,Q., Banzon,T., Lotfi,M., Davis,J., Singh,D. and Rizzolo,L.J. *et al.* (2015) Human Adult Retinal Pigment Epithelial Stem Cell-Derived RPE Monolayers Exhibit Key Physiological Characteristics of Native Tissue. *Investigative Ophthalmology & Visual Science*, **56**, 7085–7099.

17. Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

18. Li,H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, **34**, 3094–3100.

19. Wyman,D. and Mortazavi,A. (2019) TranscriptClean: Variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, **35**, 340–342.

20. Wyman,D., Balderrama-Gutierrez,G., Reese,F., Jiang,S., Rahmanian,S., Forner,S., Matheos,D., Zeng,W., Williams,B. and Trout,D. *et al.* (2020) A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*, [10.1101/672931](https://doi.org/10.1101/672931).

21. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, **29**, 15–21.

22. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–2079.

23. Pertea,G. and Pertea,M. (2020) GFF Utilities: GffRead and GffCompare. *F1000Research*, **9**, 304.

24. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, **14**, 417–419.

25. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, **20**, 110–121.

26. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J. and Thomas,S. *et al.* (2012) BEDOPS: High-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.

27. Noguchi,S., Arakawa,T., Fukuda,S., Furuno,M., Hasegawa,A., Hori,F., Ishikawa-Kato,S., Kaida,K., Kaiho,A. and Kanamori-Katayama,M. *et al.* (2017) FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*, **4**, 170112.

28. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*, **48**, D174–D179.

29. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**, 841–842.

30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

31. Lex,A., Gehlenborg,N., Strobelt,H., Vuillemot,R. and Pfister,H. (2014) UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, **20**, 1983–1992.

32. Allen,M., Poggiali,D., Whitaker,K., Marshall,T.R. and Kievit,R.A. (2019) Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, **4**, 63.

33. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biology*, **17**, 122.

34. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

35. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47–e47.

36. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : a Journal of Integrative Biology*, **16**, 284–287.

37. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

38. R Core Team (2019) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

39. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, **11**, 1650–1667.

40. Klimanskaya,I., Hipp,J., Rezai,K.A., West,M., Atala,A. and Lanza,R. (2004) Derivation and comparative assessment of retinal pigment epithelium from human embryonic stem cells using transcriptomics. *Cloning and Stem Cells*, **6**, 217–245.

41. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A. and Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Research*, **46**, D754–D761.

42. O’Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B. and Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**, D733–745.

43. Landry,J.-R., Mager,D.L. and Wilhelm,B.T. (2003) Complex controls: The role of alternative promoters in mammalian genomes. *Trends in Genetics*, **19**, 640–648.

44. Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, **18**, 18–30.

45. WANG,Y., LIU,J., HUANG,B., XU,Y.-M., LI,J., HUANG,L.-F., LIN,J., ZHANG,J., MIN,Q.-H. and YANG,W.-M. *et al.* (2015) Mechanism of alternative splicing and its regulation. *Biomedical Reports*, **3**, 152–158.

46. Takahashi,H., Kato,S., Murata,M. and Carninci,P. (2012) CAGE- Cap Analysis Gene Expression: A protocol for the detection of promoter and transcriptional networks. *Methods in molecular biology (Clifton, N.J.)*, **786**, 181–200.

47. Beck,A.H., Weng,Z., Witten,D.M., Zhu,S., Foley,J.W., Lacroute,P., Smith,C.L., Tibshirani,R., Rijn,M. van de and Sidow,A. *et al.* (2010) 3′-End Sequencing for Expression Quantification (3SEQ) from Archival Tumor Samples. *PLOS ONE*, **5**, e8768.

48. Mellough,C.B., Bauer,R., Collin,J., Dorgau,B., Zerti,D., Dolan,D.W.P., Jones,C.M., Izuogu,O.G., Yu,M. and Hallam,D. *et al.* (2019) An integrated transcriptional analysis of the developing human retina. *Development (Cambridge, England)*, **146**.

49. Gorman,S.W., Haider,N.B., Grieshammer,U., Swiderski,R.E., Kim,E., Welch,J.W., Searby,C., Leng,S., Carmi,R. and Sheffield,V.C. *et al.* (1999) The Cloning and Developmental Expression of Unconventional Myosin IXA (MYO9A) a Gene in the Bardet–Biedl Syndrome (BBS4) Region at Chromosome 15q22–q23. *Genomics*, **59**, 150–160.

50. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A. and Birnbaum,D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

51. Cummings,B.B., Karczewski,K.J., Kosmicki,J.A., Seaby,E.G., Watts,N.A., Singer-Berk,M., Mudge,J.M., Karjalainen,J., Satterstrom,F.K. and O’Donnell-Luria,A.H. *et al.* (2020) Transcript expression-aware annotation improves rare variant interpretation. *Nature*, **581**, 452–458.

52. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D. and Jang,W. *et al.* (2018) ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, **46**, D1062–D1067.

53. Reyes,A. and Huber,W. (2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, **46**, 582–592.

54. Zhang,D., Guelfi,S., Garcia-Ruiz,S., Costa,B., Reynolds,R.H., D’Sa,K., Liu,W., Courtin,T., Peterson,A. and Jaffe,A.E. *et al.* (2020) Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances*, **6**, eaay8299.

55. Yan,W., Peng,Y.-R., Zyl,T. van, Regev,A., Shekhar,K., Juric,D. and Sanes,J.R. (2020) Cell Atlas of The Human Fovea and Peripheral Retina. *Scientific Reports*, **10**, 9802.